Universidad del Valle de Guatemala

Facultad de Ingeniería

Departamento de Ciencias de la Computación

Curso: CC3084 – Data Science

Semestre II – 2025

Proyecto 1 – Obtención y Limpieza de Datos

Autores:

- José Rodrigo Marchena – 22398
- Andre Yatmian Jo – 22199
- Héctor Daniel Penedo – 22217
- Sofía Velásquez – 22049

Introducción

En este proyecto se trabaja con datos reales del Ministerio de Educación de Guatemala sobre establecimientos educativos a nivel diversificado. El objetivo principal es obtener, consolidar, limpiar y dejar listo este conjunto de datos para posteriores análisis, asegurando que el proceso de limpieza sea transparente, reproducible y bien documentado.

La limpieza se enfocará en eliminar errores comunes como inconsistencias tipográficas, duplicados, valores faltantes y variaciones no intencionales en nombres de instituciones. Las variables clave como nombre del establecimiento, dirección y teléfono recibirán especial atención, ya que son esenciales para identificar correctamente cada unidad educativa.

In [1]:
```
import pandas as pd
# Cargar el archivo CSV creado a partir del script de readfiles.py que unifica todo
file_path = "./datos/establecimiento_all.csv"
df = pd.read_csv(file_path)
```

In [2]:
```
# Mostrar dimensiones generales
print("Dimensiones iniciales del dataset:")
print(f"Número de filas: {df.shape[0]}")
print(f"Número de columnas: {df.shape[1]}")
```

```
# Visualizar primeras filas
df.head()
```

```
Dimensiones iniciales del dataset:
Número de filas: 6153
Número de columnas: 18
```

Out[2]:

	CODIGO	DISTRITO	DEPARTAMENTO	MUNICIPIO	ESTABLECIMIENTO	DIRECCION	TELE
0	16-01-0138-46	16-031	ALTA VERAPAZ	COBAN	COLEGIO COBAN	KM.2 SALIDA A SAN JUAN CHAMELCO ZONA 8	779
1	16-01-0139-46	16-031	ALTA VERAPAZ	COBAN	COLEGIO PARTICULAR MIXTO VERAPAZ	KM 209.5 ENTRADA A LA CIUDAD	773
2	16-01-0140-46	16-031	ALTA VERAPAZ	COBAN	COLEGIO "LA INMACULADA"	7A. AVENIDA 11-109 ZONA 6	782
3	16-01-0141-46	16-005	ALTA VERAPAZ	COBAN	ESCUELA NACIONAL DE CIENCIAS COMERCIALES	2A CALLE 11-10 ZONA 2	795
4	16-01-0142-46	16-005	ALTA VERAPAZ	COBAN	INSTITUTO NORMAL MIXTO DEL NORTE 'EMILIO ROSAL...	3A AVE 6-23 ZONA 11	795

Análisis inicial

El dataset cuenta con **6,153 registros** y **18 columnas**. Cada fila representa un establecimiento educativo en un municipio y departamento específico de Guatemala.
A continuación, se listan las variables que componen el dataset:

- `CODIGO` : Identificador único del establecimiento
- `DISTRITO` : Distrito escolar
- `DEPARTAMENTO` , `MUNICIPIO` : Ubicación geográfica
- `ESTABLECIMIENTO` : Nombre del establecimiento
- `DIRECCION` : Dirección física
- `TELEFONO` : Número telefónico de contacto
- `SUPERVISOR` , `DIRECTOR` : Autoridades del establecimiento
- `NIVEL` , `SECTOR` , `AREA` , `STATUS` : Clasificaciones administrativas
- `MODALIDAD` , `JORNADA` , `PLAN` : Características del modelo educativo
- `DEPARTAMENTAL` : Información adicional geográfica

- `DEPARTMENT` : Departamento de origen del archivo original

Este conjunto de datos presenta un nivel de **desorganización moderada**, con potenciales inconsistencias de formato, errores ortográficos y posibles duplicados.

In [6]:
```
# Analizar variables tipo texto con posibles inconsistencias
text_cols = df.select_dtypes(include='object')
variables_sucias = []
for col in text_cols.columns:
    nulls = df[col].isnull().sum()
    uniques = df[col].nunique()
    if nulls > 0 or uniques < len(df) * 0.95:
        variables_sucias.append(col)

variables_sucias
```

Out[6]:
```
['DISTRITO',
 'DEPARTAMENTO',
 'MUNICIPIO',
 'ESTABLECIMIENTO',
 'DIRECCION',
 'TELEFONO',
 'SUPERVISOR',
 'DIRECTOR',
 'NIVEL',
 'SECTOR',
 'AREA',
 'STATUS',
 'MODALIDAD',
 'JORNADA',
 'PLAN',
 'DEPARTAMENTAL',
 'DEPARTMENT']
```

Variables con posibles problemas de limpieza:

Se identificaron las siguientes variables con alto potencial de requerir limpieza debido a valores faltantes, inconsistencias de escritura, o cardinalidad irregular:

- `DISTRITO` , `DEPARTAMENTO` , `MUNICIPIO`
- `ESTABLECIMIENTO` , `DIRECCION` , `TELEFONO`
- `SUPERVISOR` , `DIRECTOR`
- `NIVEL` , `SECTOR` , `AREA` , `STATUS`
- `MODALIDAD` , `JORNADA` , `PLAN`
- `DEPARTAMENTAL` , `DEPARTMENT`

Estas variables probablemente presentan duplicados, uso inconsistente de mayúsculas/minúsculas, errores de ortografía, símbolos extraños o valores nulos.

Estrategias para la limpieza de datos

Siguiendo con las variables con posibles problemas de limpieza, se propone para cada una de ellas una o varias estrategias que permitan reducir la variación y mejorar los resultados del análisis.

1. DISTRITO

   Se plantea:

   - Variable categórica que agrupa a los establecimientos educativos dentro de un mismo municipio. Los primeros dos dígitos indican el departamento y los últimos tres identifican al sector. Se verificará que sean cadenas de texto que contengan solo números y el caracter "-" separando ambas partes.
   - Descartar valores faltantes o incorrectos.

2. DEPARTAMENTO

   Se plantea:

   - Se deberá verificar que todas las instancias estén escritas en **mayúsculas** y que sea un nombre existente.

3. MUNICIPIO

   Se plantea:

   - Al igual que con DEPARTAMENTO, se deberá verificar que todas las instancias estén escritas en **mayúsculas** y que el nombre de municipio exista.

4. ESTABLECIMIENTO

   Se plantea:

   - Varias de las entradas para esta columna contienen palabras encerradas en comillas ("), por lo que se propone eliminarlas para mejorar la legibilidad. También se deberá verificar que todos los nombres estén escritos en **mayúsculas**.

5. DIRECCION

   Se plantea:

   - La primera propuesta para esta variable sería, en caso de que presente demasiada variabilidad, eliminarla. Algunos establecimientos presentan direcciones únicas, por lo que no aportan demasiada información. Además, podría resultar más conveniente simplemente filtrar por municpio y departamento.
   - En caso de determinar que presenta información relevante, y al igual que con la variable ESTABLECIMIENTO, algunas de las direcciones ingresadas presentan comillas ("), por lo que también se eliminarán. Se podría extender las abreviaciones comunes, por ejemplo:

     ```
     - AVE -> AVENIDA
     - Z -> ZONA
     - COL -> COLONIA
     ```

6. TELEFONO

   Se plantea:

   - Eliminar caracteres no numéricos, validar que todos los números de teléfono contengan únicamente 8 dígitos y separar o eliminar el extra en caso de que haya varios teléfonos en el mismo campo.

7. SUPERVISOR y DIRECTOR

Se plantea:

- Ya sea eliminar u homogeneizar los títulos que aparezcan en ambas variables. Es decir, al igual que en dirección, se deberá expandir las abreviaciones comunes, por ejemplo:

```
- LIC -> LICENCIADO
- DR -> DOCTOR
- PROF -> PROFESOR
```

- Verificar que todas las instancias estén escritas en **mayúsculas** y eliminar comillas (") y espacios innecesarios.

8. NIVEL

   Se plantea:

   - Debido a que el conjunto de datos se enfoca en el nivel diversificado, todos los registros presentan el atributo DIVERSIFICADO en esta variable. Y ya que no presenta ninguna información relevante, se propone su eliminación.

9. SECTOR

   Se plantea:

   - Esta variable categórica puede tomar solamente los siguientes 3 valores "PRIVADO", "OFICIAL" o "COOPERATIVA". Se mantendran estas palabras para que sea mas sencillo identificar su significado. Asimismo, la unica limpieza que se les realizara sera asegurarnos que todas las instancias esten en **mayusculas** y esten escritas igual cuando sean la misma opcion.

10. AREA

    Se plantea:

    - Nuevamente, esta columna indica solamente dos poibilidades: area URBANA y RURAL. Debido a la simpleza de esta categoria, se mantendran las palabras como estan y la limpieza constara de asegurar que estas esten escritas de igual manera y en **mayuscula**

11. STATUS

    Se plantea:

    - **Eliminar** la columna completamente. Debido a que el origen de la data son escuelas que actualmente estan abiertas, todas las 6,153 entradas presentes estan marcadas con estatus ABIERTO. Por tanto, ya que este campo no provee informacion en lo absoluto, se removera de los datos limpiados.

12. MODALIDAD

    Se plantea:

    - Dado que las unicas opciones de esta columnas son MONOLINGUE y BILINGUE, se mantendran estas palabras y la propuesta de limpieza es asegurar que estas esten escritas adecuadamente y en **mayúsculas**

13. JORNADA

    Se plantea:

    - Mantener un estandar de escritura para las posibles opciones (VESPERTINA, MATUTINA, DOBLE y SIN JORNADA), asegurando que esten bien escritas y en mayuscula

14. PLAN

    Se plantea:

    - Para esta columna, se plantea realizar una simplificacion a las categorias presentes pues algunas son redundantes y otras un poco vagas. A continuacion se muestran las conversiones a realizar de los tipos presentes:

```
- DIARIO(REGULAR) -> REGULAR

- SEMIPRESENCIAL(DOS DIAS A LA SEMANA) ->
SEMIPRESENCIAL

- SEMIPRESENCIAL -> SEMIPRESENCIAL

- A DISTANCIA (VIRTUAL) -> VIRTUAL

- A DISTANCIA -> VIRTUAL

- FIN DE SEMANA -> FIN DE SEMANA
```

      Asimismo, se asegurara que estas opciones esten correctamente escritas y en mayusculas.

15. DEPARTAMENTAL y
16. DEPARTMENT

    Se plantea:

    - Estas dos columnas parecen contribuir cierta redundancia sobre los datos, pues en la mayoria de las observaciones, repiten su contenido. En el resto, varian ligeramente en su especificacion, por ejemplo GUATEMALA y GUATEMALA ORIENTE. De manera que se **uniran estas dos columnas** , manteniendo el registro que tenga mas informacion y asegurando que esten en mayusculas y escritas correctamente